

# Likelihood-informed dimension reduction for nonlinear inverse problems

T Cui<sup>1</sup>, J Martin<sup>2</sup>, Y M Marzouk<sup>1</sup>, A Solonen<sup>1</sup>, A Spantini<sup>1</sup>

<sup>1</sup>Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139 USA

<sup>2</sup>Institute for Computational Engineering and Sciences, The University of Texas at Austin, 201 East 24th St, Austin, TX 78712 USA

E-mail: tcui@mit.edu, jmartin@ices.utexas.edu, ymarz@mit.edu, solonen@mit.edu, spantini@mit.edu

## Abstract.

The intrinsic dimensionality of an inverse problem is affected by prior information, the accuracy and number of observations, and the smoothing properties of the forward operator. From a Bayesian perspective, changes from the prior to the posterior may, in many problems, be confined to a relatively low-dimensional subspace of the parameter space. We present a dimension reduction approach that defines and identifies such a subspace, called the “likelihood-informed subspace” (LIS), by characterizing the relative influences of the prior and the likelihood over the support of the posterior distribution. This identification enables new and more efficient computational methods for Bayesian inference with nonlinear forward models and Gaussian priors. In particular, we approximate the posterior distribution as the product of a lower-dimensional posterior defined on the LIS and the prior distribution marginalized onto the complementary subspace. Markov chain Monte Carlo sampling can then proceed in lower dimensions, with significant gains in computational efficiency. We also introduce a Rao-Blackwellization strategy that de-randomizes Monte Carlo estimates of posterior expectations for additional variance reduction. We demonstrate the efficiency of our methods using two numerical examples: inference of permeability in a groundwater system governed by an elliptic PDE, and an atmospheric remote sensing problem based on Global Ozone Monitoring System (GOMOS) observations.

Keywords: Inverse problem, Bayesian inference, dimension reduction, low-rank approximation, Markov chain Monte Carlo, variance reduction

## 1. Introduction

Inverse problems arise from indirect observations of parameters of interest. The Bayesian approach to inverse problems formalizes the characterization of these parameters through exploration of the *posterior distribution* of parameters conditioned on data [1, 2, 3]. Computing expectations with respect to the posterior distribution yields not only point estimates of the parameters (e.g., the posterior mean), but a complete description of their uncertainty via the posterior covariance and higher moments, marginal distributions, quantiles, or event probabilities. Uncertainty in parameter-dependent predictions can also be quantified by integrating over the posterior distribution.

The parameter of interest in inverse problems is often a function of space or time, and hence an element of an infinite-dimensional function space [3]. In practice, the parameter field must be discretized, and the resulting inference problem acquires a high but finite dimension. The computation of posterior expectations then proceeds via posterior sampling, most commonly using Markov chain Monte Carlo (MCMC) methods [4, 5, 6]. The computational cost and efficiency of an MCMC scheme can be strongly affected by the parameter dimension, however. The convergence rates of standard MCMC algorithms usually degrade with parameter dimension [7, 8, 9, 10, 11]; one manifestation of this degradation is an increase in the mixing time of the chain, which in turn leads to higher variance in posterior estimates. Some recent MCMC algorithms, formally derived in the infinite-dimensional setting [12, 13], do not share this scaling problem. Yet even in this setting, we will argue that significant variance reduction can be achieved through explicit dimension reduction and through de-randomization of posterior estimates, explained below.

This paper proposes a method for *dimension reduction* in Bayesian inverse problems. We reduce dimension by identifying a subspace of the parameter space that is *likelihood-informed*; this notion will be precisely defined in a relative sense, i.e., relative to the prior. Our focus is on problems with nonlinear forward operators and Gaussian priors, but builds on low-rank approximations [14] and optimality results [15] developed for the linear-Gaussian case. Our dimension reduction strategy will thus reflect the combined impact of prior smoothing, the limited accuracy or number of observations, and the smoothing properties of the forward operator. Identification of the likelihood-informed subspace (LIS) will let us write an approximate posterior distribution wherein the distribution on the complement of this subspace is taken to be independent of the data; in particular, the posterior will be approximated as the product of a low-dimensional posterior on the LIS and the marginalization of the prior onto the complement of the LIS. The key practical benefit of this approximation will be *variance reduction* in the evaluation of posterior expectations. First, Markov chain Monte Carlo sampling can be restricted to coordinates in the likelihood-informed space, enabling greater sampling efficiency—i.e., more independent samples in a given number of MCMC steps or a given computational time. Second, the product form of the approximate posterior will allow sampling in the complement of the likelihood-informed space to be avoided altogether, thus producing Rao-Blackwellized or analytically conditioned estimates of certain posterior expectations.

Dimension reduction for inverse problems has been previously pursued in several ways. [16] constructs a low dimensional representation of the parameters by using the truncated Karhunen-L  ve expansion [17, 18] of the prior distribution. A different approach, combining prior and likelihood information via low-rank approximations of

the prior-preconditioned Hessian of the log-likelihood, is used in [14] to approximate the posterior covariance in linear inverse problems. In the nonlinear setting, low-rank approximations of the prior-preconditioned Hessian are used to construct proposal distributions in the stochastic Newton MCMC method [19] or to make tractable Gaussian approximations at the posterior mode in [20]—either as a Laplace approximation, as the proposal for an independence MCMC sampler, or as the fixed preconditioner for a stochastic Newton proposal. We note that these schemes bound the tradeoff between evaluating Hessian information once (at the posterior mode) or with every sample (in local proposals). In all cases, however, MCMC sampling proceeds in the full-dimensional space.

The dimension reduction approach explored in this paper, by contrast, confines sampling to a lower-dimensional space. We extend the posterior approximation proposed in [15] to the nonlinear setting by making essentially a low-rank approximation of the *posterior expectation* of the prior-preconditioned Hessian, from which we derive a projection operator. This projection operator then yields the product-form posterior approximation discussed above, which enables variance reduction through lower-dimensional MCMC sampling and Rao-Blackwellization of posterior estimates.

We note that our dimension reduction approach does not depend on the use of any specific MCMC algorithm, or even on the use of MCMC. The low-dimensional posterior defined on coordinates of the LIS is amenable to a range of posterior exploration or integration approaches. We also note that the present analysis enables the construction of dimension-independent analogues of existing MCMC algorithms with essentially no modification. This is possible because in inverse problems with formally discretization-invariant posteriors—i.e., problems where the forward model converges under mesh refinement and the prior distribution satisfies certain regularity conditions [21, 3]—the LIS can also be discretization invariant. We will demonstrate these discretization-invariance properties numerically.

The rest of this paper is organized as follows. In Section 2, we briefly review the Bayesian formulation for inverse problems. In Section 3, we introduce the likelihood-informed dimension reduction technique, and present the posterior approximation and reduced-variance Monte Carlo estimators based on the LIS. We also present an algorithm for constructing the likelihood-informed subspace. In Section 4, we use an elliptic PDE inverse problem to demonstrate the accuracy and computational efficiency of our posterior estimates and to explore various properties of the LIS, including its dependence on the data and its discretization invariance. In Section 5, we apply our variance reduction technique to an atmospheric remote sensing problem. Section 6 offers concluding remarks.

## 2. Bayesian formulation for inverse problems

This section provides a brief overview of the Bayesian framework for the inverse problems as introduced in [1, 2, 3]. Consider the inverse problem of estimating parameters  $x$  from data  $y$ , where

$$y = G(x) + e. \quad (1)$$

Here  $e$  is a random variable representing noise and/or model error, which appears additively, and  $G$  is a known mapping from the parameters to the observables. In a Bayesian setting, we model the parameters  $x$  as a random variable and, for

simplicity, assume that the range of this random variable is a finite dimensional space  $\mathbb{X} \subseteq \mathbb{R}^n$ . Then the parameter of interest is characterized by its posterior distribution conditioned on a realization of the data,  $y \in \mathbb{Y} \subseteq \mathbb{R}^d$ :

$$\pi(x|y) \propto \pi(y|x)\pi_0(x). \quad (2)$$

We assume that all distributions have densities with respect to Lebesgue measure. The posterior probability density function above is the product of two terms: the prior density  $\pi_0(x)$ , which models knowledge of the parameters before the data are observed, and the likelihood function  $\pi(y|x)$ , which describes the probability distribution of  $y$  for any value of  $x$ .

We assume that the prior distribution is a multivariate Gaussian  $\mathcal{N}(\mu_{\text{pr}}, \Gamma_{\text{pr}})$ , where the covariance matrix  $\Gamma_{\text{pr}}$  can be also defined by its inverse,  $\Gamma_{\text{pr}}^{-1}$ , commonly referred to as the precision matrix. We model the additive noise with a zero mean Gaussian distribution, i.e.,  $e \sim \mathcal{N}(0, \Gamma_{\text{obs}})$ . This lets us define the data-misfit function

$$\eta(x) = \frac{1}{2} \left\| \Gamma_{\text{obs}}^{-\frac{1}{2}} (G(x) - y) \right\|^2, \quad (3)$$

such that the likelihood function is proportional to  $\exp(-\eta(x))$ .

### 3. Methodology

In this section, we first review the dimension reduction approach and the optimality results for linear inverse problems discussed in [14, 15]. Then we extend these ideas to nonlinear inverse problems, by constructing a likelihood-informed subspace (LIS) that captures the dominant changes from the prior to the posterior distribution. Based on the LIS, we introduce a product-form approximation of the posterior and associated reduced-variance Monte Carlo estimators. Finally, we describe an algorithm for the adaptive construction of the LIS.

#### 3.1. Optimal dimension reduction for linear inverse problems

Consider a linear forward model,  $G(x) = Gx$ , with a Gaussian likelihood and a Gaussian prior as defined in Section 2. The resulting posterior is also Gaussian,  $\pi(x|y) = \mathcal{N}(\mu_{\text{pos}}, \Gamma_{\text{pos}})$ , with mean and covariance given by

$$\mu_{\text{pos}} = \Gamma_{\text{pos}} (\Gamma_{\text{pr}}^{-1} \mu_{\text{pr}} + G^\top \Gamma_{\text{obs}}^{-1} y) \quad \text{and} \quad \Gamma_{\text{pos}} = (H + \Gamma_{\text{pr}}^{-1})^{-1}, \quad (4)$$

where

$$H = G^\top \Gamma_{\text{obs}}^{-1} G$$

is the Hessian of the data-misfit function (3). Without loss of generality we can assume zero prior mean and a positive definite prior covariance matrix.

Now consider approximations to the posterior distribution of the form

$$\tilde{\pi}(x|y) \propto \pi(y|P_r x) \pi_0(x), \quad (5)$$

where  $P_r = P_r^2$  is a rank- $r$  projector and  $\pi(y|P_r x)$  is an approximation to the original likelihood function  $\pi(y|x)$ . Approximations of this form can be computationally advantageous when operations involving the prior (e.g., evaluations or sampling) are less expensive than those involving the likelihood. As described in [15], they are also the natural form with which to approximate a Bayesian update, particularly in the inverse problem setting with high-dimensional  $x$ . In the deterministic case, inverse

problems are ill-posed; the data cannot inform all directions in the parameter space. Equivalently, the spectrum of  $H$  is compact or decays quickly. Thus one should be able to explicitly project the argument of the likelihood function onto a low-dimensional space without losing much information in the process. The posterior covariance remains full rank, but the update from prior covariance to posterior covariance will be low rank. The challenge, of course, is to find the best projector  $P_r$  for any given  $r$ . The answer will involve balancing the influence of the prior and the likelihood. In the following theorem, we introduce the optimal projector and characterize its approximation properties.

**Theorem 1.** *Let  $\Gamma_{\text{pr}} = LL^\top$  be a symmetric decomposition of the prior covariance matrix and let  $(\lambda_i, v_i)$  be the eigenvalue-eigenvector pairs of the prior-preconditioned Hessian  $(L^\top HL)$  such that  $\lambda_i \geq \lambda_{i+1}$ . Define the directions  $u_i = Lv_i$  and  $w_i = L^{-\top}v_i$  together with the matrices  $U_r = [u_1, \dots, u_r]$  and  $W_r = [w_1, \dots, w_r]$ . Then, the projector  $P_r$  given by:*

$$P_r = U_r W_r^\top,$$

*yields an approximate posterior density of the form  $\tilde{\pi}(x|y) = \mathcal{N}(\mu_{\text{pos}}^{(r)}, \Gamma_{\text{pos}}^{(r)})$  and is optimal in the following sense:*

- (i)  $\Gamma_{\text{pos}}^{(r)}$  minimizes the Förstner distance [22] from the exact posterior covariance over the class of positive definite matrices that can be written as rank  $r$  negative semidefinite updates of the prior covariance.
- (ii)  $\mu_{\text{pos}}^{(r)} = A^*y$  minimizes the Bayes risk  $\mathbb{E}_{x,y} [\|\mu(y) - x\|_{\Gamma_{\text{pos}}^{-1}}^2]$  over the class of all linear transformations of the data  $\mu(y) = Ay$  with  $\text{rank}(A) \leq r$ .

*Proof.* We refer the reader to [15] for a proof and detailed discussion.  $\square$

The vectors  $(u_1, \dots, u_r)$  span the range of the optimal projector; we call this range the *likelihood-informed subspace* of the linear inverse problem. Note that the  $(u_i)$  are eigenvectors of the pencil  $(H, \Gamma_{\text{pr}}^{-1})$ . Hence, the  $j$ th basis vector  $u_j$  maximizes the Rayleigh quotient

$$\mathcal{R}(u) = \frac{\langle u, Hu \rangle}{\langle u, \Gamma_{\text{pr}}^{-1}u \rangle} \quad (6)$$

over the subspace  $\mathbb{X} \setminus \text{span}\{u_1, \dots, u_{j-1}\}$ . This Rayleigh quotient helps interpret the  $(u_i)$  as directions where the data are most “informative” relative to the prior. For example, consider a direction  $w \in \mathbb{X}$  representing a rough mode in the parameter space. If the prior is smoothing, then the denominator of (6) will be large; also, if the forward model output is relatively insensitive to variation in the  $w$  direction, the numerator of (6) will be small. Thus the Rayleigh quotient will be small and  $w$  is not particularly data-informed relative to the prior. Conversely, if  $w$  is smooth then the prior variance in this direction may be large and the likelihood may be relatively constraining; this direction is then data-informed. Of course, there are countless intermediate cases, but in general, directions for which (6) are large will lie in the range of  $U_r$ .

Note also that  $U_r$  diagonalizes both  $H$  and  $\Gamma_{\text{pr}}^{-1}$ . We are particularly interested in the latter property: the modes  $(u_i)$  are orthogonal (and can be chosen orthonormal) with respect to the inner product induced by the prior precision matrix. This property

will be preserved later in the nonlinear case, and will be important to our posterior sampling schemes.

For nonlinear inverse problems, we seek an approximation of the posterior distribution in the same form as (5). In particular, the range of the projector will be determined by blending together likelihood-informed subspaces from regions of high posterior probability. The construction of the approximation will be detailed in the following section. Clearly, the optimality statements in Theorem 1 do not hold in a nonlinear setting. However, they can still be used to construct optimal approximations to *local* likelihood-informed subspaces, as we will see shortly.

### 3.2. LIS construction for nonlinear inverse problems

When the forward model is nonlinear, the Hessian of the data-misfit function varies over the parameter space, and thus the likelihood-informed directions are embedded in some nonlinear manifold. We aim to construct a global linear subspace to capture the majority of this nonlinear likelihood-informed manifold.

Let the forward model  $G(x)$  be first-order differentiable. The linearization of the forward model at a given parameter value  $x$ ,  $J(x) \equiv \nabla G(x)$  where  $J(x) \in \mathbb{R}^{d \times n}$ , provides the local sensitivity of the parameter-to-observable map. Inspired by the dimension reduction approach for the linear inverse problem, we use the linearized forward model  $J(x)$  to construct the Gauss-Newton approximation of the Hessian of the data-misfit function,

$$H(x) = J(x)^\top \Gamma_{\text{obs}}^{-1} J(x).$$

Now consider a local version of the Rayleigh quotient (6),

$$\mathcal{R}(u; x) := \frac{\langle u, H(x)u \rangle}{\langle u, \Gamma_{\text{pr}}^{-1}u \rangle}.$$

Introducing the change of variable  $v = L^{-1}u$ , we can equivalently use

$$\tilde{\mathcal{R}}(v; x) := \frac{\langle v, (L^\top H(x) L)v \rangle}{\langle v, v \rangle} = \mathcal{R}(Lv; x), \quad (7)$$

to quantify the local impact of the likelihood relative to the prior. As in the linear problem, this suggests the following procedure for computing a local LIS given some truncation threshold  $\tau_{\text{loc}}$ :

**Problem 2** (Construction of the local likelihood-informed subspace). *Given the Gauss-Newton Hessian of the data misfit function  $H(x)$  at a given  $x$ , find the eigendecomposition of the prior-preconditioned Gauss-Newton Hessian (ppGNH)*

$$L^\top H(x) L v_i = \lambda_i v_i. \quad (8)$$

*Given a truncation threshold  $\tau_{\text{loc}} > 0$ , the local LIS is spanned by  $U_l = [u_1, \dots, u_l]$ , where  $u_i = L v_i$  corresponds to the  $l$  leading eigenvalues such that  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_l \geq \tau_{\text{loc}}$ .*

For a direction  $u$  with  $\mathcal{R}(u; x) = 1$ , the local impact of the likelihood and the prior are balanced. Thus, to retain a comprehensive set of likelihood-informed directions, we typically choose a truncation threshold  $\tau_{\text{loc}}$  less than 1.

To extend the pointwise criterion (7) into a global criterion for likelihood-informed directions, we consider the expectation of the Rayleigh quotient over the posterior

$$\mathbb{E}_\pi [\mathcal{R}(u; x)] = \mathbb{E}_\pi [\tilde{\mathcal{R}}(v; x)] = \frac{\langle v, Sv \rangle}{\langle v, v \rangle},$$

where  $S$  is the expected ppGNH over the posterior,

$$S = \int_{\mathbb{X}} L^\top H(x) L \pi(dx|y). \quad (9)$$

Then we can naturally construct the global LIS through the eigendecomposition of  $S$  as in the linear case. We consider approximating  $S$  using the Monte Carlo estimator

$$\hat{S}_n = \frac{1}{n} \sum_{k=1}^n L^\top H(x^{(k)}) L,$$

where  $x^{(k)} \sim \pi(x|y)$ ,  $k = 1 \dots n$ , are posterior samples. Since the local Hessian  $H(x^{(k)})$  is usually not explicitly available and is not feasible to store for large-scale problems, we use its prior-preconditioned low-rank approximation as defined in Problem 2. Thus the global LIS can be constructed by the following procedure:

**Problem 3** (Construction of global likelihood-informed subspace). *Suppose we have a set of posterior samples  $\mathcal{X} = \{x^{(k)}\}$ ,  $k = 1 \dots m$ , where for each sample  $x^{(k)}$ , the ppGNH is approximated by the truncated low rank eigendecomposition*

$$L^\top H(x^{(k)}) L \approx \sum_{i=1}^{l(k)} \lambda_i^{(k)} v_i^{(k)} v_i^{(k)\top},$$

*by solving Problem 2. We have  $\lambda_i^{(k)} \geq \tau_{loc}$  for all  $k = 1 \dots m$  and all  $i = 1 \dots l(k)$ . To construct the global LIS, we consider the eigendecomposition of the Monte Carlo estimator of the expected Hessian in (9), which takes the form*

$$\left( \frac{1}{m} \sum_{k=1}^m \sum_{i=1}^{l(k)} \lambda_i^{(k)} v_i^{(k)} v_i^{(k)\top} \right) \psi_i = \gamma_i \psi_i. \quad (10)$$

*The global LIS has the non-orthogonal basis  $\Phi_r = L\Psi_r$ , where the eigenvectors  $\Psi_r = [\psi_1, \dots, \psi_r]$  correspond to the  $r$  leading eigenvalues of (10),  $\gamma_1 \geq \dots \geq \gamma_r \geq \tau_g$ , for some truncation threshold  $\tau_g > 0$ . Here we choose  $\tau_g$  to be equal to the threshold  $\tau_{loc}$  in Problem 2.*

In many applications we can only access the Gauss-Newton Hessian by computing its action on vectors, which involves one forward model evaluation and one adjoint model evaluation. In such a case, the ppGNH can be approximated by finding the eigendecomposition (8) using either Krylov subspace algorithms [23] or randomized algorithms [24, 25].

We have not yet discussed how to obtain the posterior samples used in Problem 3 and how many samples are required. Depending on how strongly  $H(x)$  (or its dominant eigenspace) varies over the support of the posterior, the number of samples required to construct the global LIS may vary significantly. To automate this process and fully explore the directional landscape of the likelihood, we adaptively enrich the LIS by adding samples to (10) until the change in the LIS falls below a certain threshold. The details are described in Section 3.5.

### 3.3. Posterior approximation

Our goal is to make a computationally efficient approximation of the posterior distribution by projecting the likelihood function onto the subspace that is informed by the data, i.e., the LIS. Such an approximate posterior has the form

$$\tilde{\pi}(x|y) \propto \pi(y|\Pi_r x) \pi_0(x), \quad (11)$$

where  $\Pi_r$  is a projector onto the LIS. We choose a projector that is self-adjoint with respect to the inner product induced by the prior precision matrix. This projector  $\Pi_r$  induces a natural decomposition of the parameter space as  $\mathbb{X} = \mathbb{X}_r \oplus \mathbb{X}_\perp$ , where  $\mathbb{X}_r = \text{range}(\Pi_r)$  is the LIS and  $\mathbb{X}_\perp = \text{range}(I - \Pi_r)$  is the complement subspace (CS). This choice leads to a factorization of the prior distribution into the product of two distributions, one defined on the low-dimensional LIS and the other on the CS. This factorization is the key to our dimension reduction technique.

**Definition 4.** We define the projectors  $\Pi_r$  and  $I - \Pi_r$ , and a corresponding parameter decomposition, as follows:

- (a) Suppose the LIS basis computed in Problem 3 is  $\Phi_r = L\Psi_r$ , where  $\Psi_r$  is orthonormal. Define the matrix  $\Xi_r = L^{-\top}\Psi_r$  such that  $\Xi_r^\top \Phi_r = I_r$ . The projector  $\Pi_r$  has the form

$$\Pi_r = \Phi_r \Xi_r^\top.$$

Choose  $\Psi_\perp$  such that  $[\Psi_r \ \Psi_\perp]$  forms a complete orthonormal system in  $\mathbb{R}^n$ . Then the projector  $I - \Pi_r$  can be written as

$$I - \Pi_r = \Phi_\perp \Xi_\perp^\top,$$

where  $\Phi_\perp = L\Psi_\perp$  and  $\Xi_\perp = L^{-\top}\Psi_\perp$ .

- (b) Naturally, the parameter  $x$  can be decomposed as

$$x = \Pi_r x + (I - \Pi_r)x,$$

where each projection can be represented as the linear combination of the corresponding basis vectors. Consider the “LIS parameter”  $x_r$  and the “CS parameter”  $x_\perp$ , which are the weights associated with the LIS basis  $\Phi_r$  and CS basis  $\Phi_\perp$ , respectively. Then we can define the following pair of linear transformations between the parameter  $x$  and  $(x_r, x_\perp)$ :

$$x = [\Phi_r \ \Phi_\perp] \begin{bmatrix} x_r \\ x_\perp \end{bmatrix}, \quad \begin{bmatrix} x_r \\ x_\perp \end{bmatrix} = [\Xi_r \ \Xi_\perp]^\top x. \quad (12)$$

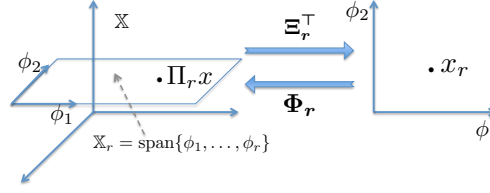
Figure 1 illustrates the transformations between the parameter projected onto the LIS,  $\Pi_r x$ , and the LIS parameter  $x_r$ . The same relation holds for the transformations between  $(I - \Pi_r)x$  and the CS parameter  $x_\perp$ . And as Definition 4 makes clear,  $\Pi_r$  is an oblique projector.

**Lemma 5.** Suppose we have  $x = \Phi_r x_r + \Phi_\perp x_\perp$  as defined in Definition 4(b). Then the prior distribution can be decomposed as

$$\pi_0(x) = \pi_r(x_r) \pi_\perp(x_\perp),$$

where  $\pi_r(x_r) = \mathcal{N}(\Xi_r^\top \mu_{\text{pr}}, I_r)$  and  $\pi_\perp(x_\perp) = \mathcal{N}(\Xi_\perp^\top \mu_{\text{pr}}, I_\perp)$ .





**Figure 1.** Illustration of the transformation between the parameter projected onto the LIS,  $\Pi_r x$ , and the LIS parameter  $x_r$ .

Following Definition 4(b) and Lemma 5, the approximate posterior distribution (11) can be reformulated as

$$\begin{aligned}\tilde{\pi}(x|y) &\propto \pi(y|\Pi_r x) \pi_r(x_r) \pi_\perp(x_\perp) \\ &= \pi(y|\Phi_r x_r) \pi_r(x_r) \pi_\perp(x_\perp).\end{aligned}$$

Applying the linear transformation from  $x$  to  $(x_r, x_\perp)$  as defined in Equation (12), we can rewrite the approximate posterior for the parameters  $(x_r, x_\perp)$  as

$$\tilde{\pi}(x_r, x_\perp|y) \propto \tilde{\pi}(x|y) \propto \tilde{\pi}(x_r|y) \pi_\perp(x_\perp), \quad (13)$$

which is the product of the reduced posterior

$$\tilde{\pi}(x_r|y) \propto \pi(y|\Phi_r x_r) \pi_r(x_r), \quad (14)$$

and the complement prior  $\pi_\perp(x_\perp)$ . To compute a Monte Carlo estimate of the expectation of a function over the approximate posterior distribution (13), we only need to sample the reduced posterior  $\tilde{\pi}(x_r|y)$ , since properties of the Gaussian complement prior  $\pi_\perp(x_\perp)$  are known analytically.

One can also combine MCMC samples from the reduced posterior  $\tilde{\pi}(x_r|y)$  with independent samples from the complement prior  $\pi_\perp(x_\perp)$  to provide samples that are approximately drawn from the full posterior  $\pi(x|y)$ . By correcting these samples via importance weights or a Metropolis scheme, one would then obtain a sampling algorithm for the original full-space posterior. This idea is not pursued further here, and in the rest of this work we will emphasize the analytical properties of the complement prior  $\pi_\perp(x_\perp)$ , using them to reduce the variance of Monte Carlo estimators.

### 3.4. Reduced-variance estimators

Suppose we have a function  $h(x)$  for which the conditional expectation over the approximate posterior (11)

$$\mathbb{E}_{\tilde{\pi}}[h(x)|x_r] = \int_{\mathbb{X}_\perp} h(\Phi_r x_r + \Phi_\perp x_\perp) \pi_0(x_\perp) dx_\perp, \quad (15)$$

can be calculated either analytically or through some high-precision numerical quadrature scheme. Then variance reduction can be achieved as follows:

- (i) **Subspace MCMC.** Use MCMC in the LIS to simulate a “subspace Markov chain” with target distribution  $\tilde{\pi}(x_r|y)$  (13). Any number of MCMC algorithms developed in the literature can be applied off-the-shelf, e.g., adaptive MCMC [26, 27, 28, 29, 30], the stochastic Newton algorithm of [19], or the Riemannian manifold algorithms of [31]. Since the dimension of the LIS can be quite small

relative to the original parameter space, the subspace MCMC approach can yield lower sample correlations (better mixing) than applying any of these MCMC algorithms directly to the full posterior (2).

- (ii) **Rao-Blackwellization.** We approximate  $\mathbb{E}_\pi[h] = \int_{\mathbb{X}} h(x)\pi(dx|y)$  by the expectation of the function  $h(x)$  over the approximate posterior  $\tilde{\pi}(x|y)$ , i.e.,  $\mathbb{E}_{\tilde{\pi}}[h] = \int_{\mathbb{X}} h(x)\tilde{\pi}(dx|y)$ . Given a set of subspace MCMC samples  $\{x_r^{(1)}, \dots, x_r^{(N)}\}$  where  $x_r^{(k)} \sim \tilde{\pi}_r(x_r|y)$ , a Monte Carlo estimator of  $\mathbb{E}_{\tilde{\pi}}[h]$  is given by

$$\tilde{Q}_N = \frac{1}{N} \sum_{k=1}^N \mathbb{E}_{\tilde{\pi}} [h(x)|x_r^{(k)}]. \quad (16)$$

As an application of the Rao-Blackwellization principle (see [32] and references therein), the estimator (16) has a lower variance than the standard estimator

$$Q_N = \frac{1}{N} \sum_{k=1}^N h(x^{(k)}), \quad (17)$$

where  $x^{(k)} \sim \pi(x|y)$ .

This procedure mitigates many of the difficulties of posterior exploration in high dimensions, provided that the prior-to-posterior update is reasonably low rank. Variance reduction is achieved not only by increasing the effective sample size per MCMC iteration (via subspace MCMC), but also by reducing the variance of Monte Carlo estimators using Rao-Blackwellization. In effect, we argue that the Gaussian CS can be explored *separately* and via a calculation (15) that does not involve sampling.

Note that while this procedure necessarily reduces the variance of a Monte Carlo estimator, it introduces bias since we replace the expectation over the full posterior  $\mathbb{E}_\pi[h]$  with an expectation over the approximate posterior  $\mathbb{E}_{\tilde{\pi}}[h]$ . Thus this variance reduction is particularly useful in situations where the variance of the estimator (17) derived from full-space MCMC samples is large compared with the bias, which is often the case for high-dimensional inverse problems.

Beyond variance reduction, subspace MCMC offers several additional computational advantages over MCMC methods applied to the full posterior directly: (i) The storage requirement for saving subspace MCMC samples is much lower than that of an MCMC scheme that samples the full posterior. (ii) For MCMC methods where the proposal distribution involves operations with square root of the prior covariance matrix (e.g., the stochastic Newton [19] and preconditioned Crank-Nicolson [12, 13] techniques) the computational cost of handling the full prior covariance can be much higher than the computational cost of handling the reduced prior  $\pi_r(x_r)$ , which has identity covariance.

The Monte Carlo estimator (16) can be further simplified if the function of interest  $h(x)$  can be expressed as either the product or the sum of two separate functions,  $h_r(x_r)$  and  $h_\perp(x_\perp)$ , defined on the LIS and CS, respectively. In the multiplicative case  $h(x) = h_r(x_r)h_\perp(x_\perp)$ , the conditional expectation (15) can be written as

$$\mathbb{E}_{\tilde{\pi}} [h(x)|x_r] = h_r(x_r) \int_{\mathbb{X}_\perp} h_\perp(x_\perp)\pi_0(dx_\perp).$$

In the additive case  $h(x) = h_r(x_r) + h_\perp(x_\perp)$ , it can be written as

$$\mathbb{E}_{\tilde{\pi}} [h(x)|x_r] = h_r(x_r) + \int_{\mathbb{X}_\perp} h_\perp(x_\perp)\pi_0(dx_\perp).$$

Thus the expectation  $\mathbb{E}_{\tilde{\pi}}[h]$  can be decomposed either into the product (in the multiplicative case) or the sum (in the additive case) of the pair of expectations

$$\mathbb{E}_{\tilde{\pi}}[h_r] = \int_{\mathbb{X}_r} h_r(x_r) \pi(dx_r|y), \quad (18)$$

$$\mathbb{E}_{\tilde{\pi}}[h_{\perp}] = \int_{\mathbb{X}_{\perp}} h_{\perp}(x_{\perp}) \pi_0(dx_{\perp}), \quad (19)$$

which are associated with the LIS and CS, respectively. The expectation in (18) can be computed by the subspace MCMC methods described above, whereas the expectation in (19) is computed analytically or through high-order numerical integration.

Now we give two particularly useful examples of the analytical treatment of the complement space.

**Example 6** (Reduced variance estimation of the posterior mean). *Suppose we have obtained the empirical posterior mean  $\tilde{\mu}_r$  of the reduced parameter  $x_r$  using subspace MCMC. The resulting reduced-variance estimator of the posterior mean is*

$$\mathbb{E}_{\tilde{\pi}}[x] = \Phi_r \tilde{\mu}_r + \Pi_{\perp} \mu_{\text{pr}} = \Phi_r \tilde{\mu}_r + (I - \Pi_r) \mu_{\text{pr}}.$$

**Example 7** (Reduced variance estimator of the posterior covariance). *Suppose we have the empirical posterior covariance  $\tilde{\Gamma}_r$  of the reduced parameter  $x_r$ , estimated using subspace MCMC. The resulting reduced-variance estimator of the posterior covariance is*

$$\begin{aligned} \text{Cov}_{\tilde{\pi}}[x] &= \Phi_r \tilde{\Gamma}_r \Phi_r^{\top} + \Pi_{\perp} \Gamma_{\text{pr}} \Pi_{\perp}^{\top} \\ &= \Gamma_{\text{pr}} + \Phi_r^{\top} \left( \tilde{\Gamma}_r - I_r \right) \Phi_r. \end{aligned}$$

### 3.5. Algorithms for the LIS

Constructing the global LIS requires a set of posterior samples. Since the computational cost of solving Problem 2 for any sample is much greater than the cost of evaluating the forward model, we wish to limit the number of samples used in Problem 3 while ensuring that we adequately capture the posterior variation of the ppGNH. Thus we choose samples using the following adaptive procedure.

**Algorithm 8** (Global LIS construction using subspace MCMC). *First, compute the posterior mode  $x_{\text{map}} \in \mathbb{X}$ . Set the initial sample set for Problem 3 to  $\mathcal{X}^{(1)} = \{x_{\text{map}}\}$ . Solve Problem 3 to find  $\Psi_r^{(1)}$ , the initial LIS basis  $\Phi_r^{(1)}$ , and its left-inverse  $\Xi_r^{(1)}$ .<sup>‡</sup> Initialize a subspace Markov chain with initial state  $\Xi_r^{(1)\top} x_{\text{map}}$ , which is the posterior mode projected onto the LIS. At any subsequent step  $k \geq 1$ , the following procedure is used to adaptively enrich the LIS:*

- (i) **Subchain simulation.** *Simulate the  $r(k)$ -dimensional subspace MCMC chain for  $L$  iterations, so that the last state of this chain, denoted by  $\theta$ , is uncorrelated with its initial state. Then  $\theta$  transformed back to the original parameter space,  $(\Phi_r^{(k)} \theta)$ , is used as the next sample point. Enrich the sample set to  $\mathcal{X}^{(k+1)} = \mathcal{X}^{(k)} \cup \{\Phi_r^{(k)} \theta\}$ .*

<sup>‡</sup> The dimension of the global LIS can vary at each iteration. Let  $r(k)$  denote the dimension of the global LIS at iteration  $k$ . To be precise, we should then write  $\Phi_{r(k)}^{(k)}$  and  $\Xi_{r(k)}^{(k)}$ , but for brevity we will simplify notation to  $\Phi_r^{(k)}$  and  $\Xi_r^{(k)}$  when possible.

- (ii) **LIS construction.** Solve Problem 3 with the sample set  $\mathcal{X}^{(k+1)}$ . Then update the LIS basis to  $\Phi_r^{(k+1)}$  and  $\Xi_r^{(k+1)}$ . Set the initial state of the next subspace MCMC chain to  $\Xi_r^{(k+1)\top} \Phi_r^{(k)} \theta$ .
- (iii) **Convergence checking.** Terminate the adaptation if a pre-specified maximum allowable number of Hessian evaluations is exceeded, or if the weighted subspace distance in Definition 9 falls below a certain threshold. Otherwise, set  $k \leftarrow k + 1$  and return to Step (i).

The convergence criterion in step (iii) is based on an incremental distance between likelihood-informed subspaces. The distance penalizes changes in the dominant directions (those with large eigenvalues  $\gamma$ ) more heavily than changes in the less important directions (those with smaller  $\gamma$ ).

**Definition 9** (Weighted subspace distance). At iteration  $k$ , define the basis/weights pair  $\mathcal{Y}^{(k)} = \{\Psi_r^{(k)}, D^{(k)}\}$ , where  $\Psi_r^{(k)}$  is the orthonormal LIS basis from Problem 3 and  $D_{ij}^{(k)} = \delta_{ij} \hat{\gamma}_i^{(k)}$  is a diagonal matrix consisting of normalized weights

$$\hat{\gamma}_i^{(k)} = \frac{\gamma_i^{(k)}}{\sum_{j=1}^{r(k)} \gamma_j^{(k)}}, \quad j = 1 \dots r(k),$$

computed from the eigenvalues  $\{\gamma_1^{(k)}, \dots, \gamma_{r(k)}^{(k)}\}$  of Problem 3. For two adjacent steps  $k$  and  $k + 1$ , we compute the weighted subspace distance of [33], which has the form

$$\mathcal{D}(\mathcal{Y}^{(k)}, \mathcal{Y}^{(k+1)}) = \sqrt{1 - \left\| \left( \Psi_{r(k)}^{(k)} (D^{(k)})^{\frac{1}{4}} \right)^\top \left( \Psi_{r(k+1)}^{(k+1)} (D^{(k+1)})^{\frac{1}{4}} \right) \right\|_F^2}. \quad (20)$$

Note that in Step (i) of Algorithm 8, we construct the global LIS by always sampling in an adaptively enriched subspace. This offers computational benefits, since the MCMC exploration is always confined to a lower dimensional space. However, a potential problem with this approach is that it might ignore some directions that are also data-informed. A more conservative approach would be to introduce a *conditional update* at the end of each subchain simulation: perform Metropolized independence sampling in the current CS using the complement prior as proposal. This would enable the subchain to explore the full posterior, but would result in higher-dimensional sampling when constructing the LIS. In our numerical examples, described below, no conditional updates were required for good performance; constructing the LIS using samples from the full posterior and using the subspace approach gave essentially the same results. Of course, one could also simply employ a standard MCMC algorithm to sample the full posterior, and then construct the LIS using the resulting posterior samples. However, the efficiency of the MCMC algorithm in this case will be affected by the dimensionality of the problem.

#### 4. Example 1: Elliptic PDE

Our first example is an elliptic PDE inverse problem, e.g., inferring the permeability field of a two dimensional stationary groundwater equation from partial observations of the pressure field. We use this test case to demonstrate several aspects of our dimension reduction technique: (i) construction of the LIS, and the impact of mesh refinement on the structure of the LIS; (ii) application of low-rank posterior mean and variance estimators to the permeability field; (iii) changes in the LIS with varying amounts of observational data.

#### 4.1. Problem setup

Consider the problem domain  $\Omega = [0, 3] \times [0, 1]$ , with boundary  $\partial\Omega$ . We denote the spatial coordinate by  $s \in \Omega$ . Consider the permeability field  $\kappa(s)$ , the pressure field  $p(s)$ , and sink/source terms  $f(s)$ . The pressure field for a given permeability and source/sink configuration is governed by the Poisson equation

$$\begin{cases} -\nabla \cdot (\kappa(s) \nabla p(s)) &= f(s), & s \in \Omega \\ \langle \kappa(s) \nabla p(s), \vec{n}(s) \rangle &= 0, & s \in \partial\Omega \end{cases} \quad (21)$$

where  $\vec{n}(s)$  is the outward normal vector on the boundary. To make a well-posed boundary value problem, a further boundary condition

$$\int_{\partial\Omega} p(s) ds = 0, \quad (22)$$

is imposed. The source/sink term  $f(s)$  is defined by the superposition of four weighted Gaussian plumes with standard deviation (i.e., spatial width) 0.05, centered at four corners  $[0, 0]$ ,  $[3, 0]$ ,  $[3, 1]$  and  $[0, 1]$ , with weights  $\{1, 2, 3, -6\}$ . The system of equations (21) is solved by the finite element method with  $120 \times 40$  bilinear elements.

The discretized permeability field  $\kappa$  is endowed with a log-normal prior distribution, i.e.,

$$\kappa = \exp(x), \text{ and } x \sim \mathcal{N}(0, \Gamma_{\text{pr}}), \quad (23)$$

where the covariance matrix  $\Gamma_{\text{pr}}$  is defined through an anisotropic exponential covariance kernel

$$\text{Cov}(x(s), x(s')) = \sigma_u^2 \exp \left( -\frac{((s - s')^\top \Sigma^{-1} (s - s'))^{\frac{1}{2}}}{s_0} \right), \text{ for } s, s' \in \Omega. \quad (24)$$

In this example, we set the anisotropic correlation tensor to

$$\Sigma = \begin{bmatrix} 0.55 & -0.45 \\ -0.45 & 0.55 \end{bmatrix},$$

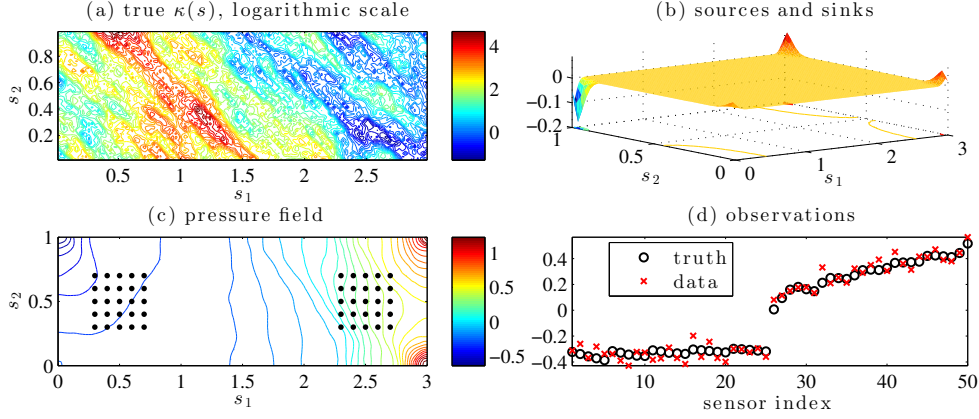
the prior standard deviation to  $\sigma_u = 1.15$ , and the correlation length to  $s_0 = 0.18$ . The “true” permeability field is a realization from the prior distribution. The true permeability field, the sources/sinks, the simulated pressure field, and the synthetic data are shown in Figure 2.

Partial observations of the pressure field are collected at 50 measurement sensors as shown by the black dots in Figure 2(c). The observation operator  $M$  is simply the corresponding “mask” operation. This yields observed data  $y \in \mathbb{R}^{50}$  as

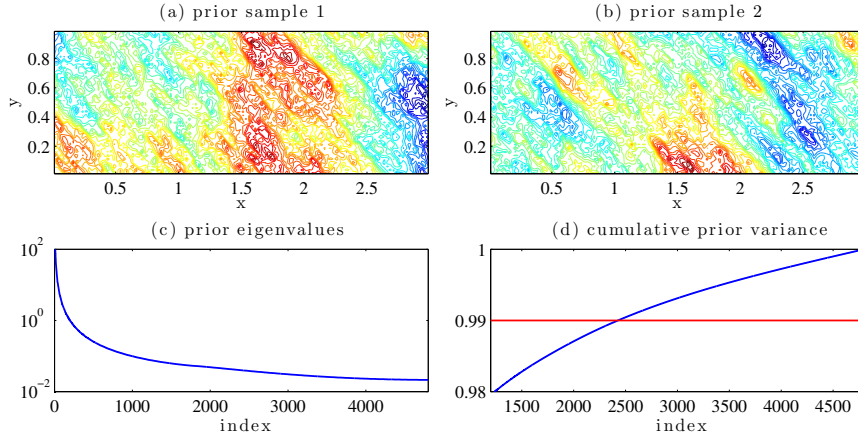
$$y = Mp(s) + e,$$

with additive error  $e \sim \mathcal{N}(0, \sigma^2 I_{50})$ . The standard deviation  $\sigma$  of the measurement noise is prescribed so that the observations have signal-to-noise ratio 10, where the signal-to-noise ratio is defined as  $\max_s \{p(s)\} / \sigma$ . The noisy data are shown in Figure 2(d).

Figure 3 shows two draws from the prior distribution, the eigenspectrum of the prior covariance, and the cumulative prior variance integrated over  $\Omega$  (i.e., the running sum of the prior covariance eigenvalues). In order to keep 99% percent of the energy in the prior, 2427 eigenmodes are required. Because of this slow decay of the prior covariance spectrum, *a priori* dimension reduction based on a truncated eigendecomposition of the prior covariance (as described in [16]) would be very inefficient for this problem. Information carried in high-frequency eigenfunctions cannot be captured unless an enormous number of prior modes are retained; thus, a better basis is required.



**Figure 2.** Setup of the elliptic inversion example. (a) “True” permeability field. (b) Sources and sinks. (c) Pressure field resulting from the true permeability field, with measurement sensors indicated by black circles. (d) Data  $y$ ; circles represent the noise-free pressure at each sensor, while crosses represent the pressure observations corrupted with measurement noise.

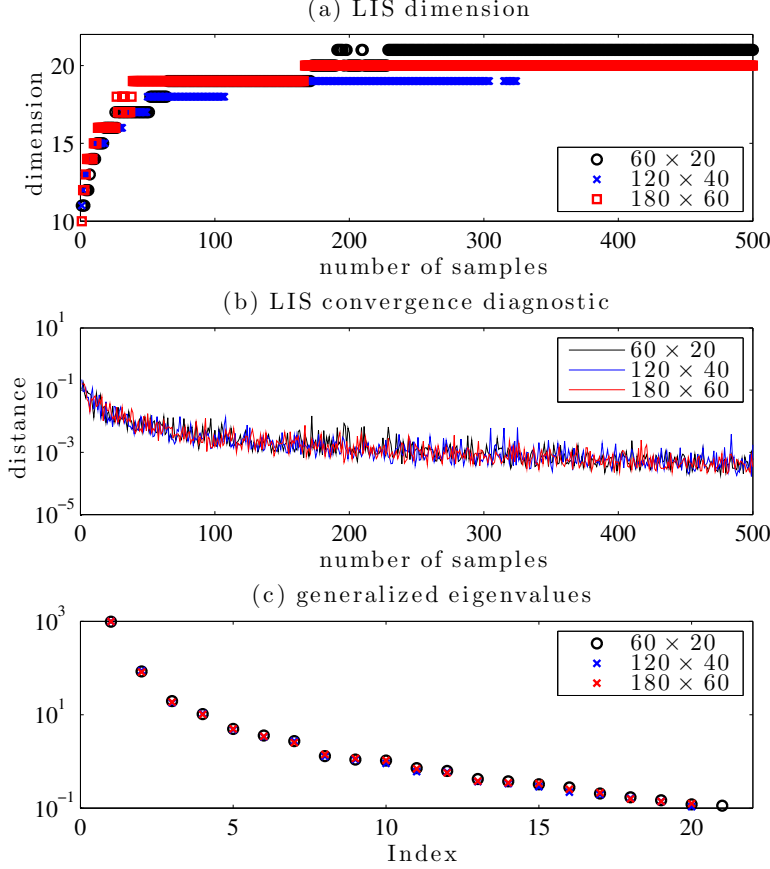


**Figure 3.** Prior samples and eigenspectrum of the prior covariance. (a) and (b): Two samples drawn from the prior. (c) Prior covariance spectrum, eigenvalues versus index number. (d) Cumulative energy (integrated prior variance) over a subset of the eigenspectrum, shown in blue; the red line represents the 99% energy truncation threshold.

#### 4.2. LIS construction

Now we demonstrate the process of LIS construction using Algorithm 8, and the structure of the LIS under mesh refinement. To compute the LIS, we run Algorithm 8 for 500 iterations, using adaptive MALA [28] to simulate each subchain with length  $L = 200$ . We choose the truncation thresholds  $\tau_{loc} = \tau_g = 0.1$ . To explore the dimensionality and structure of the LIS versus mesh refinement, we carry out the same numerical experiment on a  $60 \times 20$  coarse grid, a  $120 \times 40$  intermediate grid, and a  $180 \times 60$  fine grid. The dimension of the LIS versus number of iterations, the

evolution of the convergence diagnostic (20), and the generalized eigenvalues after 500 iterations—for each level of grid refinement—are shown in Figure 4. Also, Figure 5 shows the first five LIS basis vectors for each level of discretization.

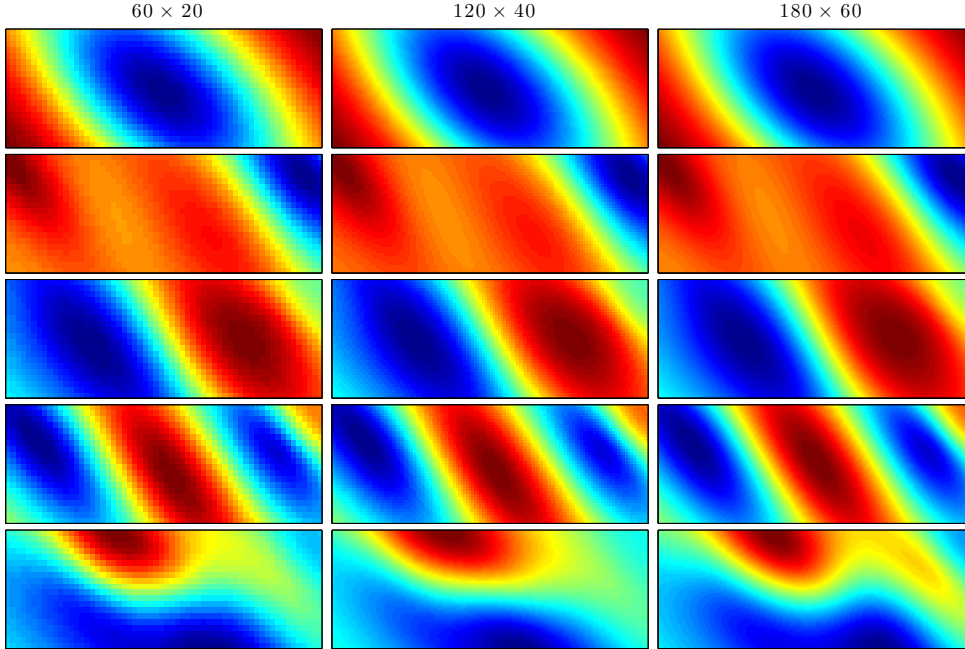


**Figure 4.** The dimension of the LIS and the convergence diagnostic (20) versus the number of samples used in the adaptive construction. Black, blue, and red markers represent the  $60 \times 20$  grid, the  $120 \times 40$  grid, and the  $180 \times 60$  grid, respectively. Subplot (a) shows the dimension of the LIS; subplot (b) shows the weighted distance between successive subspaces; and subplot (c) shows the generalized eigenvalues  $\gamma_i^{(k)}$  after  $k = 500$  iterations.

As shown in Figure 4(a), the dimension of the LIS changes rapidly in the first 100 iterations, then it stabilizes. Change in the dimension reflects the fact that the log-likelihood Hessian  $H(x)$  varies locally in this non-Gaussian problem. We also observe that the  $60 \times 20$  grid has a slightly larger final LIS dimension than the two refined grids: at the end of the adaptive construction, the LIS of the  $60 \times 20$  grid has dimension 21, while the  $120 \times 40$  grid and the  $180 \times 60$  grid yield LIS dimensions of 20. This effect may be ascribed to larger discretization errors in the  $60 \times 20$  grid.

The weighted distance (20) between each adjacent pair of likelihood-informed subspaces is used as the convergence diagnostic during the construction process. With any of the three discretizations, the weighted subspace distance at the end of adaptive





**Figure 5.** The first five LIS basis vectors (columns of  $\Phi_5$ ) for different levels of discretization of the inversion parameters  $x$ . In the figure, columns 1–3 correspond to the  $60 \times 20$  grid, the  $120 \times 40$  grid, and the  $180 \times 60$  grid, respectively. The basis vectors in each column are ordered top to bottom by decreasing eigenvalue.

construction is several orders of magnitude lower than at the beginning, as shown in Figure 4(b). We also observe that the rates of convergence of this diagnostic are comparable for all three levels of discretization. These figures suggest that while local variation of the Hessian is important in this problem (e.g., the dimension of the LIS doubles over the course of the iterations), much of this variation is well-explored after 100 or 200 iterations of Algorithm 8.

Since the forward model converges with grid refinement, we expect that the associated LIS should also converge. The generalized eigenvalues for all three grids are shown in Figure 4(c), where the spectra associated with all three subspaces have very similar values. And as shown in Figure 5, the leading LIS basis vectors  $\{\varphi_1, \dots, \varphi_5\}$  have similar shapes for all three levels of grid refinement. Refinement leads to slightly more structure in  $\varphi_5$ , but the overall mode shapes are very close.

#### 4.3. Estimation of the posterior mean and variance

With an LIS in hand, we apply the variance reduction procedure described in Section 3.3 to estimate the posterior mean and variance of the permeability field. Calculations in this subsection use the  $120 \times 40$  discretization of the PDE and inversion parameters.

We first demonstrate the sampling performance of subspace MCMC, where we use adaptive MALA [28] to sample the LIS-defined reduced posterior  $\tilde{\pi}(x_r|y)$  (14). We compare the results of subspace MCMC with the results of Hessian-preconditioned Langevin MCMC applied to the full posterior  $\pi(x|y)$  (2) (referred to as “full-space MCMC” hereafter). The latter MCMC scheme results from an explicit discretization of



the Langevin SDE, preconditioned by the inverse of the log-posterior Hessian evaluated at the posterior mode (see [34] for details). Note that we cannot precondition the full-dimensional Langevin SDE by the empirical posterior covariance as in adaptive MALA because of the high parameter dimension ( $n = 4800$ ). In this setup, subspace MCMC and full-space MCMC require the same number of forward model and gradient evaluations for a given number of MCMC iterations.

To examine sampling performance, the autocorrelation of the log-likelihood function and the autocorrelations of the parameters projected onto the first, third, and fifth LIS basis vectors are used as benchmarks. These results are shown in Figure 6. We run both algorithms for  $10^6$  iterations and discard the first half of the chains as burn-in. The top row of Figure 6 shows these benchmarks for both samplers. For all four benchmarks, subspace MCMC produces a faster decay of autocorrelation as a function of sample lag—i.e., a lower correlation between samples after any given number of MCMC steps.

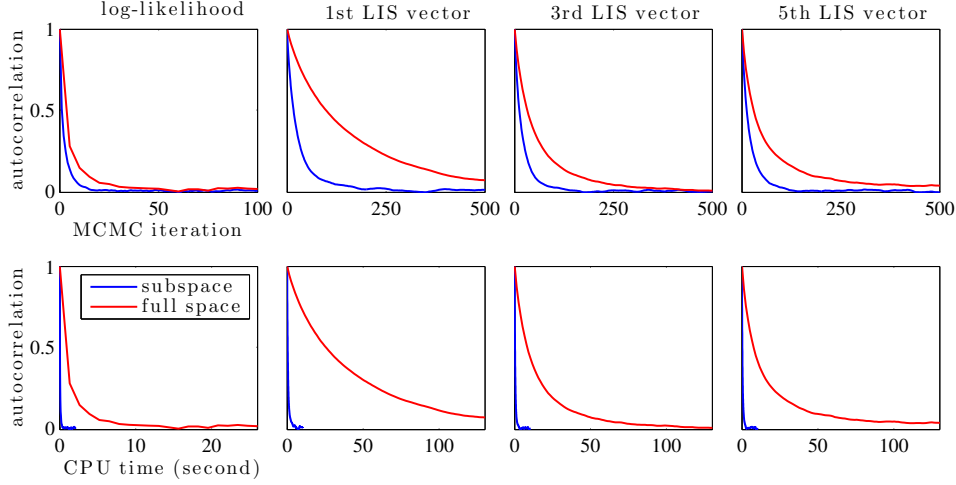
Furthermore, as discussed in Section 3.3, even though the same number of forward model evaluations are required by subspace MCMC and full-space MCMC for a given number of samples, the computational cost of operations involving the square root of the prior covariance—used in sampling and evaluating the proposal distribution—can be much higher for full-space MCMC than subspace MCMC. In this test case, running subspace MCMC for  $10^6$  iterations cost  $2.1 \times 10^4$  seconds of CPU time, while running full-space MCMC for the same number of iterations took  $2.6 \times 10^5$  seconds. To incorporate this cost difference, the second row of Figure 6 shows the autocorrelation of the four benchmark quantities as a function of CPU time rather than sample lag. Here, we immediately observe that the autocorrelation per CPU time is further reduced by using subspace MCMC.

Of course, recall that to construct the LIS we simulated Algorithm 8 for 500 iterations. This costs roughly  $2.2 \times 10^4$  seconds of CPU time, which is only 8.5% of the time required to run full-space MCMC for  $10^6$  steps. Therefore subspace MCMC, including the cost of LIS construction, takes less time to produce a given number of samples than full-space MCMC *and* these samples are less correlated—i.e., of higher quality.

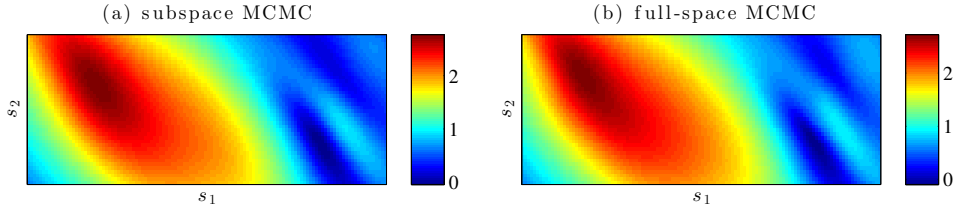
We now compare reduced-variance estimates of the posterior mean and variance (obtained with subspace MCMC) with estimates obtained via full-space MCMC. The results are shown in Figures 7 and 8. Full-space MCMC and subspace MCMC yield very similar mean and variance estimates. Figures 8(a) and (b) distinguish the two components of the Rao-Blackwellized variance estimates described in Example 7. Variance in the LIS, shown in Figure 8(a), is estimated from MCMC samples, while variance in the CS, shown in Figure 8(b), is calculated analytically from the prior and the LIS projector. The sum of these two variance fields is shown in Figure 8(c), and it is nearly the same as the full-space result in Figure 8(d). In the central part of the domain where measurement sensors are not installed, we can observe that the variance is larger in the CS than in the LIS, and hence this part of the domain is prior-dominated. In the right part of the domain, the variance is less prior-dominated, since this region is covered by observations.

#### 4.4. The influence of data

The amount of information carried in the data affects the dimension and structure of the LIS. To demonstrate the impact of the data, we design a case study where different

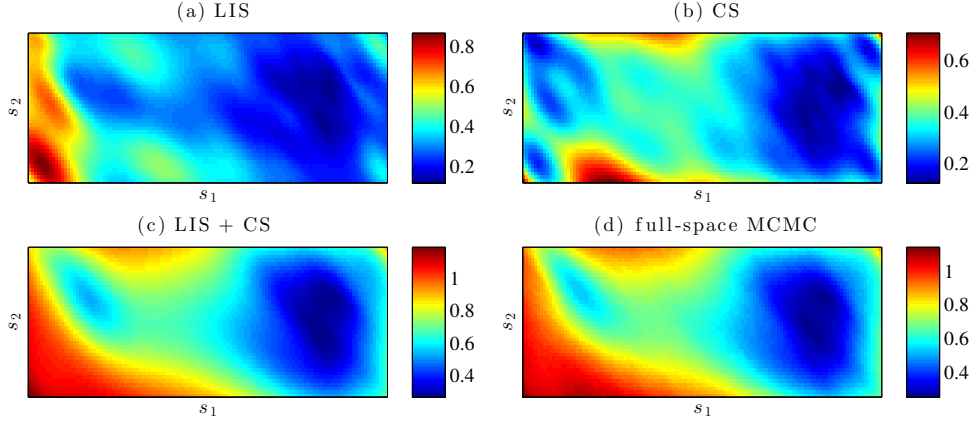


**Figure 6.** Autocorrelations of various benchmarks: blue line is subspace MCMC and red line is full-space MCMC. Column 1: log-likelihood function. Column 2: parameters projected onto the first LIS basis vector. Column 3: parameters projected onto the third LIS basis vector. Column 4: parameters projected onto the fifth LIS basis vector. Top row: Autocorrelation as a function of sample lag. Bottom row: Autocorrelation as a function of sample lag, where the latter is measured via CPU time.

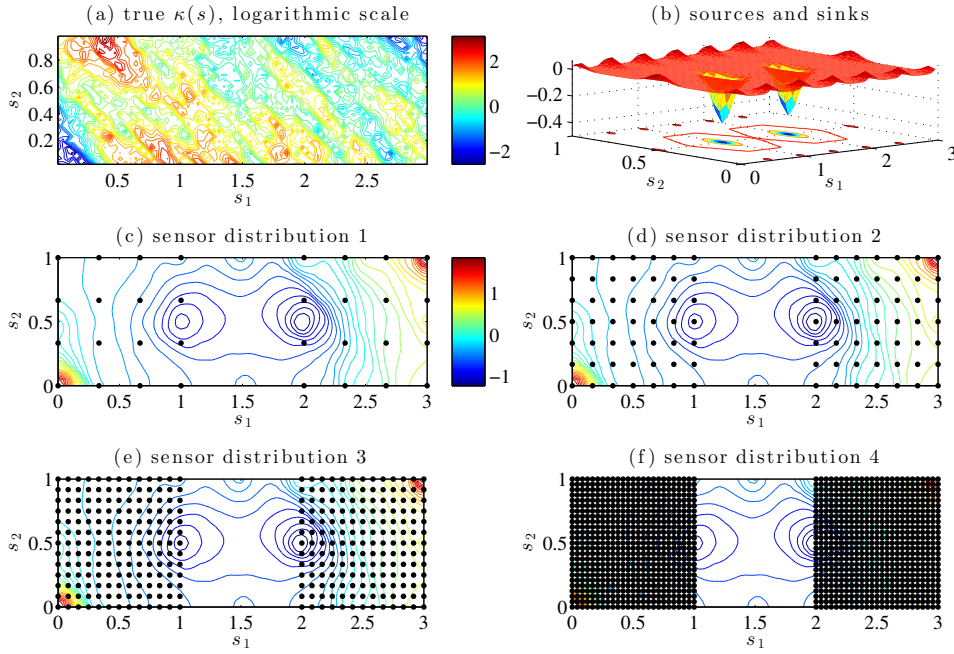


**Figure 7.** Estimates of posterior mean: (a) using subspace MCMC, (b) using full-space MCMC.

likelihood-informed subspaces are constructed under various observational scenarios. The same stationary groundwater problem defined in Section 4.1 is employed here. For the sake of computational efficiency, the problem domain  $\Omega = [0, 3] \times [0, 1]$  is discretized by a slightly coarser  $72 \times 24$  mesh. And to provide a stronger impulse to the groundwater system, the source/sink terms used in this example are different from those used in Sections 4.1–4.3. Along the boundary of the domain  $\Omega$ , we evenly distribute a set of sources with a distance of 0.5 between the source centers. Two sinks are placed in the interior of the domain at locations  $[0.5, 1]$  and  $[2, 0.5]$ . Each source has weight 1, while each sink has weight 3.5. We distributed sensors evenly over the domain  $[0, 1] \times [0, 1] \cup [2, 3] \times [0, 1]$ ; starting with an inter-sensor spacing of  $1/3$ , we incrementally refine the sensor distribution with spacings of  $1/6$ ,  $1/12$ , and  $1/24$ . This results in four different data sets, containing the noisy readings of 32, 98, 338, and 1250 sensors, respectively. The true permeability field, the sources/sinks, the simulated pressure field, and sensor distributions are shown in Figure 9.



**Figure 8.** Estimation of the posterior variance: (a) empirical estimate using MCMC in the LIS; (b) analytical evaluation in the CS; (c) combined LIS + CS estimate; (d) for comparison, estimation using full-space MCMC.

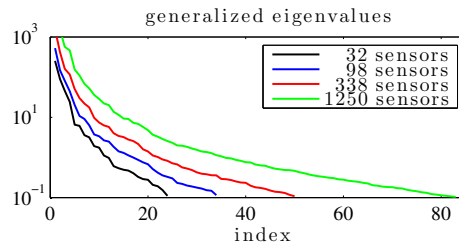


**Figure 9.** Setup of the elliptic inversion example for testing the influence of data. (a) True permeability field. (b) Sources and sinks. (c)–(f) Pressure field resulting from the permeability field defined in (a), and sensor distributions (black dots) for data sets 1–4.

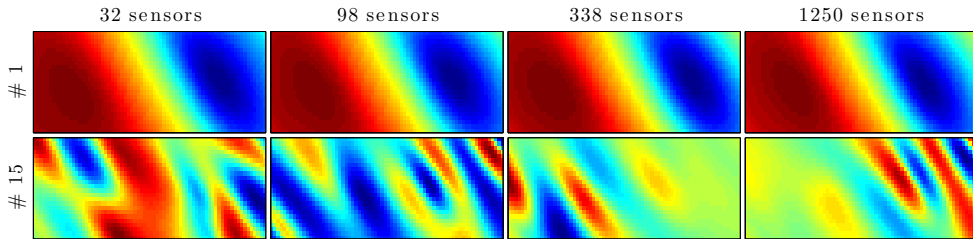
As in Section 4.2, we run Algorithm 8 for 500 iterations to construct the LIS, using subchains of length  $L = 200$ . For data sets 1–4, the resulting LISs have dimension 24, 34, 50, and 83, respectively. The generalized eigenvalue spectrum for each LIS is

shown in Figure 10. We note that the eigenvalues decay more slowly with increasing amounts of data. This behavior is expected; since the generalized eigenvalues reflect the impact of the likelihood, relative to the prior, more data should lead to more directions where the likelihood dominates the prior.

Since the sensors for all four data sets occupy the same area of the spatial domain, we expect that the four likelihood-informed subspaces should share a similar low frequency structure. However, the high frequency structures in each LIS might differ from each other under refinement of the sensor distribution. Thus the LIS basis vectors corresponding to the largest eigenvalues should share a similar pattern, while the LIS basis vectors corresponding to the relatively small eigenvalues might have different patterns. We observe this effect in the numerical experiments carried out here; Figure 11 shows the first and fifteenth LIS basis vector for each of the data sets.



**Figure 10.** Generalized eigenvalues associated with the likelihood-informed subspace under refinement of the observations. The black, blue, red, and green lines show eigenvalues for data sets 1–4, with 32, 98, 338, and 1250 sensors, respectively.



**Figure 11.** The first and fifteenth LIS basis vectors for each of the four data sets.

## 5. Example 2: atmospheric remote sensing

In this section, we apply the dimension reduction approach to a realistic atmospheric satellite remote sensing problem. The problem is to invert the concentrations of various gases in the atmosphere using the measurement system applied in the GOMOS satellite instrument, which stands for *Global Ozone MONitoring System*.

GOMOS is an instrument on board ESA’s Envisat satellite, and was operational for about 10 years before the connection with the satellite was lost in May 2012. The GOMOS instrument performs so-called star occultation measurements; it measures, at different wavelengths, the absorption of starlight as it travels through the atmosphere.

Different gases in the atmosphere (such as ozone, nitrogen dioxide and aerosols) leave fingerprints in the measured intensity spectra. The task of the inversion algorithm is to infer the concentrations of these gases based on the measurements.

The GOMOS inverse problem is known to be ill-posed; the intensity spectra may contain strong information about the major gases (like  $O_3$ ) at some altitudes, whereas some minor gases (like aerosols) at some altitudes may be practically unidentifiable and totally described by the prior. Thus, the GOMOS problem is a good candidate for our approach: the dimension of the likelihood informed subspace is expected to be small and the prior contribution large.

Next, we briefly present the GOMOS theory and the inverse problem setup. For more details about the GOMOS instrument and the Bayesian treatment of the inverse problem, see [35] and the references therein.

### 5.1. The GOMOS model

The GOMOS instrument repeatedly measures light intensities  $I_\lambda$  at different wavelengths  $\lambda$ . First, a reference intensity spectrum  $I_{\text{ref}}$  is measured above the atmosphere. The so-called transmission spectrum is defined as  $T_\lambda = I_\lambda/I_{\text{ref}}$ . The transmissions measured at wavelength  $\lambda$  along the ray path  $l$  are modelled using Beer's law:

$$T_{\lambda,l} = \exp \left( - \int_l \sum_{\text{gas}} \alpha_\lambda^{\text{gas}}(z(s)) \rho^{\text{gas}}(z(s)) ds \right), \quad (25)$$

where  $\rho^{\text{gas}}(z(s))$  is the density of a gas (unknown parameter) at tangential height  $z$ . The so called cross-sections  $\alpha_\lambda^{\text{gas}}$ , known from laboratory measurements, define how much a gas absorbs light at a given wavelength.

To approximate the integrals in (25), the atmosphere is discretized. The geometry used for inversion resembles an onion: the gas densities are assumed to be constant within spherical layers around the Earth. The GOMOS measurement principle is illustrated in Figure 12 below.

Here, we assume that the cross-sections do not depend on height. In the inverse problem we have  $n_{\text{gas}}$  gases,  $n_\lambda$  wavelengths, and the atmosphere is divided into  $n_{\text{alts}}$  layers. The discretisation is fixed so that number of measurement lines is equal to the number of layers. Approximating the integrals by sums in the chosen grid, and combining information from all lines and all wavelengths, we can write the model in matrix form as follows:

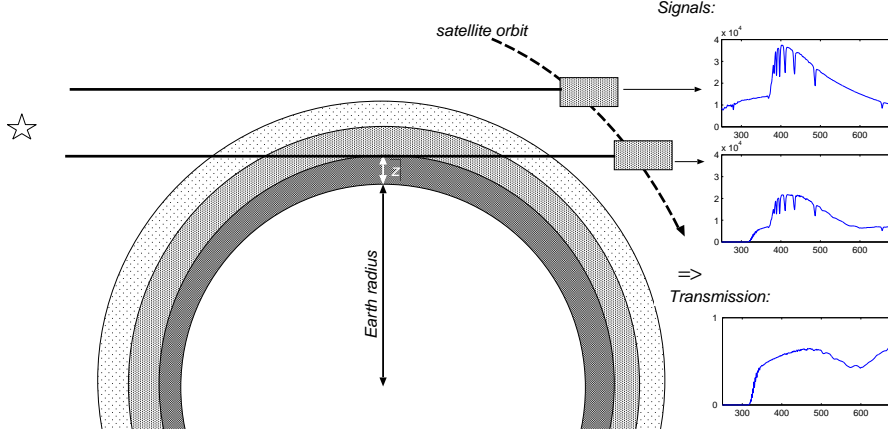
$$T = \exp(-CB^\top A^\top), \quad (26)$$

where  $T \in \mathbb{R}^{n_\lambda \times n_{\text{alts}}}$  are the modelled transmissions,  $C \in \mathbb{R}^{n_\lambda \times n_{\text{gas}}}$  contains the cross-sections,  $B \in \mathbb{R}^{n_{\text{alts}} \times n_{\text{gas}}}$  contains the unknown densities and  $A \in \mathbb{R}^{n_{\text{alts}} \times n_{\text{alts}}}$  is the geometry matrix that contains the lengths of the lines of sight in each layer.

Computationally, it is convenient to deal with vectors of unknowns. We vectorize the above model using the identity  $\text{vec}(CB^\top A^\top) = (A \otimes C)\text{vec}(B^\top)$ , where  $\otimes$  denotes the Kronecker product and  $\text{vec}$  is the standard vectorization obtained by stacking the columns of the matrix argument on top of each other. Thus, the likelihood model is written in vector form as follows:

$$y = \text{vec}(T) + e = \exp(-(A \otimes C)\text{vec}(B^\top)) + e, \quad (27)$$

where  $e$  is the measurement error, for which we apply an independent Gaussian model with known variances.



**Figure 12.** The principle of the GOMOS measurement. The reference intensity is measured above the atmosphere. The observed transmission spectrum is the attenuated spectrum (measured through the atmosphere) divided by the reference spectrum. The atmosphere is presented locally as spherical layers around the Earth. Note that the thickness of the layers is much larger relative to the Earth in this figure than in reality. The figure is adopted from [35], with the permission of the authors.

Note that, in principle, the model (27) could be linearized by taking logarithms of both sides, which is usually done for such tomography problems (like X-ray computerized tomography). For this problem, linearisation can cause problems, since the signal from the star is often smaller compared to the background noise in the measurement.

### 5.2. Data and prior

Here, we generate synthetic data by solving the forward model (27) with known gas densities  $x$ . In the example, we have 4 gas profiles to be inverted. The atmosphere is discretized into 50 layers, and the total dimension of the problem is thus 200. The simulated data are illustrated in Figure 13.

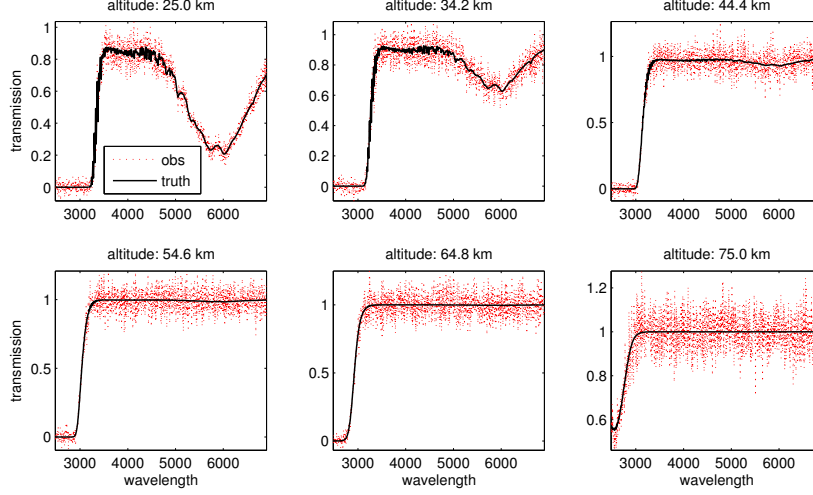
We estimate the log-profiles  $x = \log(\text{vec}(B^\top))$  of the gases instead of the densities  $B$  directly. We set a Gaussian process prior for the profiles, which yields  $x_i \sim N(\mu_i, \Sigma_i)$ , where  $x_i$  denotes the elements of vector  $x$  corresponding to gas  $i$ . The elements of the  $50 \times 50$  covariance matrices are calculated based on the squared exponential covariance function

$$C_i(s, s') = \sigma_i \exp(-(s - s')^2 / 2s_{0,i}^2), \quad (28)$$

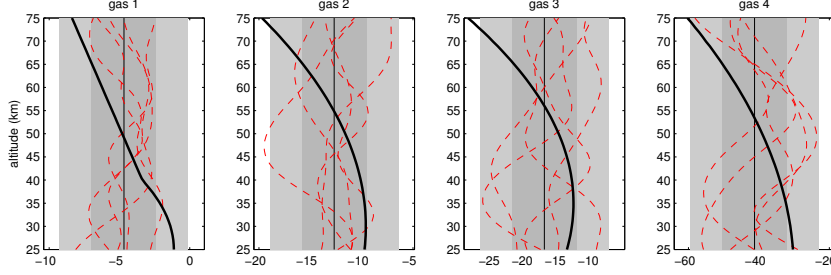
where the parameter values are  $\sigma_1 = 5.22$ ,  $\sigma_2 = 9.79$ ,  $\sigma_3 = 23.66$ ,  $\sigma_4 = 83.18$ , and  $s_{0,i} = 10$  for all  $i$ . The priors are chosen to promote smooth profiles and to give a rough idea about the magnitude of the density values. The prior is illustrated in Figure 14.

### 5.3. Inversion results

In this particular synthetic example, we know that gas 1 is very well identified by the data. The data also contain information about gases 2 and 3 at some altitudes. Gas



**Figure 13.** GOMOS example setup: the true transmissions (black) and the observed transmissions (red) for 6 altitudes.

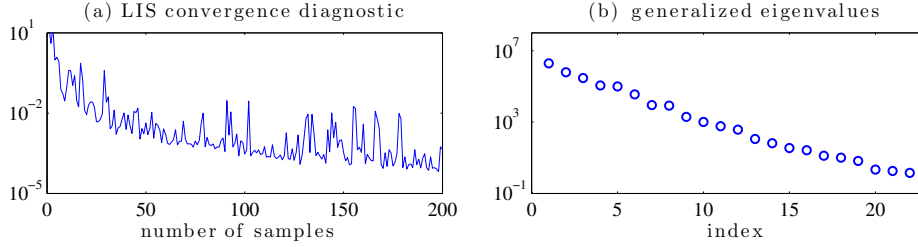


**Figure 14.** True log-profiles for the 4 gases (black solid lines), 50% and 95% confidence envelopes for the prior (grey areas) and 5 samples from the prior (red dashed lines).

4, on the other hand, is totally unidentified by the data.

The LIS is constructed using 200 samples—i.e., 200 iterations of Algorithm 8—starting with the Hessian at the posterior mode. The subspace convergence diagnostic and the generalized eigenvalues are shown in Figure 15. We choose the truncation thresholds  $\tau_{loc} = \tau_g = 0.5$ . The dimension of the LIS in the end was 22.

We compute  $10^6$  samples in both the LIS and in the full 200-dimensional space using the Hessian-preconditioned MALA algorithm. In Figure 16, the first two columns show the mean gas profile and the mean  $\pm 1$  and 2 standard deviations in the LIS and in the complement space (CS). The third column shows the combined posterior from the LIS and the CS; for comparison, results from full-space MCMC are shown in the fourth column. Note the different scales on the horizontal axes throughout the figure. We observe that the subspace approach, where MCMC is applied only in a 22-dimensional space, yields results very similar to those of full MCMC. In addition, comparing the contributions of the LIS and CS indicates that gas 1 is dominated by



**Figure 15.** Left: the convergence diagnostic (20) versus the number of samples used to construct the LIS. Right: the generalized eigenvalues associated with the final LIS.

the likelihood, whereas the posterior distribution of gas 4 is entirely determined by the prior. Note that the CS contribution for gas 1 is tiny (check the scale), while the LIS contribution for gas 4 is also very small. For gases 2 and 3, the lower altitudes are likelihood-dominated, while the higher altitudes have more contribution from the prior. The full-space MCMC results for gas 4 show a slightly non-uniform mean, but this appears to be the result of sampling variance. By avoiding sampling altogether in the CS, the subspace approach most likely yields a more accurate posterior in this case.

To further illustrate the approach, we plot the first six basis vectors of the LIS in Figure 17. One can see that the first basis vectors mainly include features of gas 1, which is most informed by the data. The first basis vectors also contain some features of gases 2 and 3 in lower altitudes. Gas 4 is not included in the LIS at all.

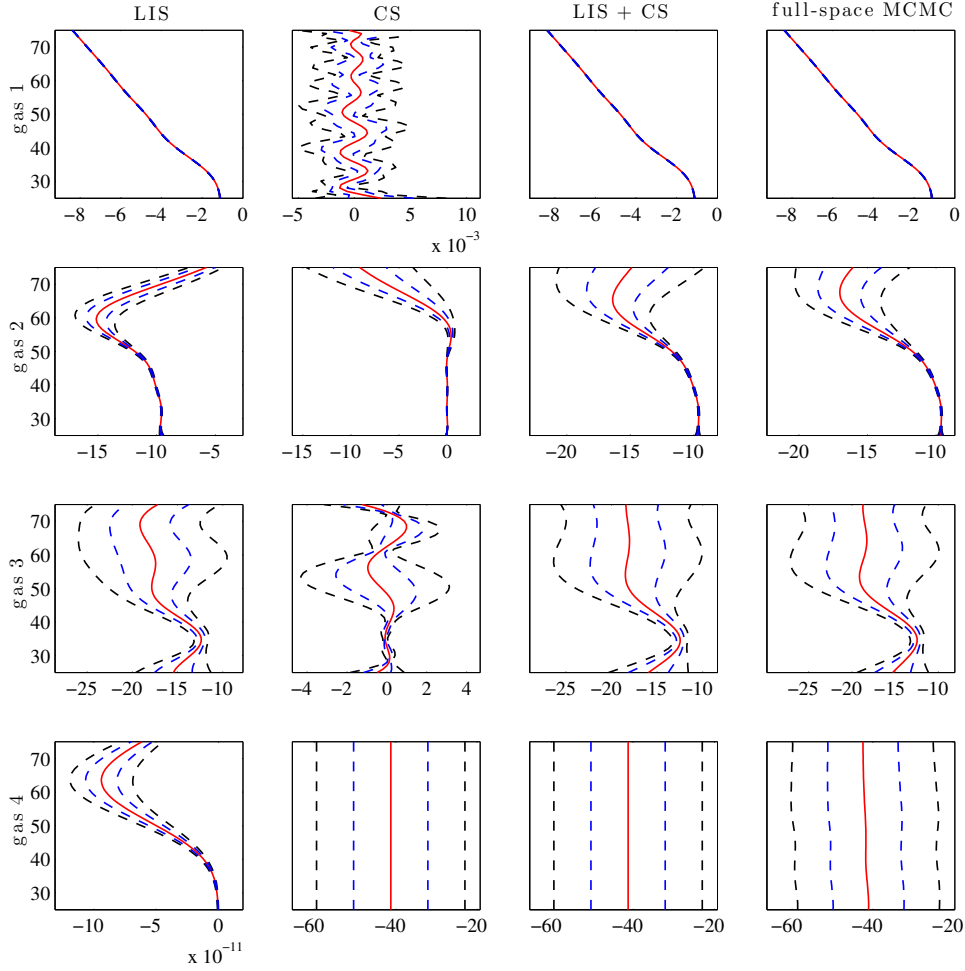
The dimension reduction obtained via the subspace approach is expected to yield better mixing than the full-space MCMC. For the GOMOS case, the chain autocorrelations for subspace and full-space MCMC are compared in Figure 18. The subspace sampler shows much faster decay of the autocorrelations than full-space MCMC.

In this test case, the subspace MCMC also has lower computational cost compared to full-space MCMC. To simulate a Markov chain for  $10^6$  iterations, the subspace MCMC consumed about 2560 seconds of CPU time, while the full-space MCMC cost 3160 CPU seconds. We note that the CPU time reduction is not as significant as the elliptic example, because the prior covariance is a  $200 \times 200$  dimensional matrix, which is much smaller than the covariance matrix used in the elliptic example. To construct the LIS, we simulated Algorithm 8 for 200 iterations. This cost about 136 seconds of CPU time, which is only about 4.3% of the CPU time used to run full-space MCMC for  $10^6$  steps.

## 6. Conclusions

In this paper, we present a new approach for dimension reduction in nonlinear inverse problems with Gaussian priors. Our approach is based on dividing the parameter space into two subspaces: a likelihood-informed subspace (LIS) where the likelihood has a much greater influence on the posterior than the prior distribution, and the complement to the LIS where the Gaussian prior dominates. We explore the posterior projected onto the LIS (the “difficult” and non-Gaussian part of the problem) with Markov chain Monte Carlo while treating the complement space as exactly Gaussian.

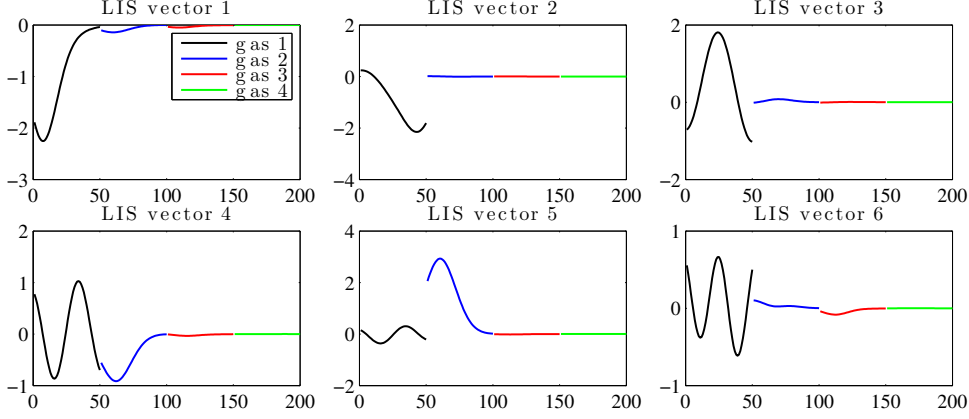




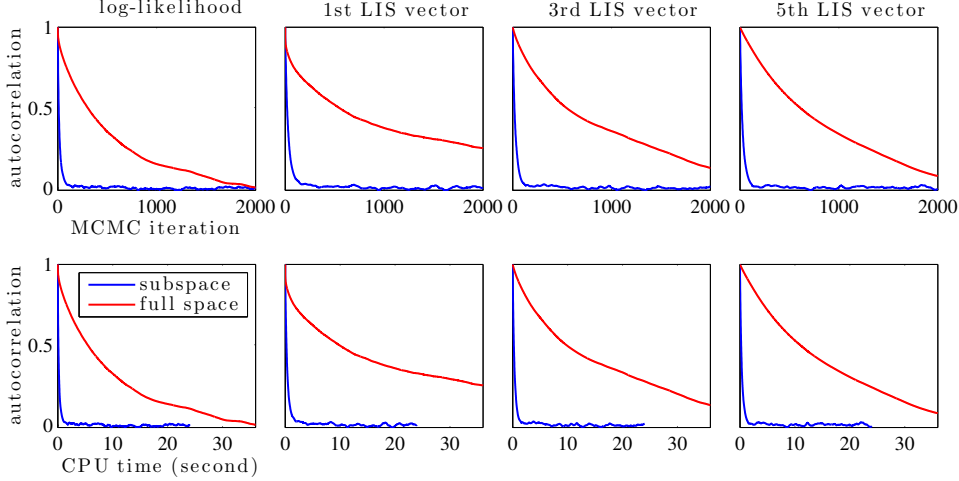
**Figure 16.** Mean and  $\pm 1/\pm 2$  standard deviations for the 4 gas profiles computed from the LIS samples alone (1st column), CS alone (2nd column) and when the LIS and CS are combined (3rd column). The same quantities computed from full-space MCMC are given in the 4th column.

This approximation allows us to analytically integrate many functions over the complement space when estimating their posterior expectations; the result is a Rao-Blackwellization or de-randomization procedure that can greatly reduce the variance of posterior estimates. Particularly in inverse problems—where information in the data is often limited and the solution of the problem relies heavily on priors—the dimension of the LIS is expected to be small, and the majority of the directions in the parameter space can be handled analytically.

The dimension reduction approach is based on theory developed for the linear case; in [15] it is shown that in linear-Gaussian problems, the eigendecomposition of the prior-preconditioned log-likelihood Hessian yields an optimal low-rank update from the prior to the posterior, which can be interpreted in terms of a projector whose range is the LIS. Here, we generalize the approach to nonlinear problems,



**Figure 17.** The first six LIS basis vectors for the remote sensing example. The colors indicate the components of the unknown vector corresponding to the different gases. In each subfigure, the  $x$ -axis denotes the index of the parameter vector, and, for each gas, the components are ordered from low altitudes to high altitudes. (For example, the black line in each figure shows gas 1 profiles from low altitudes to high altitudes, etc.)



**Figure 18.** Autocorrelations of full-space (red) and subspace (blue) MCMC for the log-likelihood (1st column) and for the samples projected onto the first, third, and fifth LIS basis vectors (2nd, 3rd and 4th columns). Top rows shows the autocorrelations computed per MCMC step and bottom row per CPU time.

where the log-likelihood Hessian varies over the parameter space. Our solution is to construct many local likelihood-informed subspaces over the support of the posterior and to combine them into a single global LIS. We show how the global LIS can be constructed efficiently in an adaptive manner, starting with the LIS computed at the posterior mode and iteratively enriching the global LIS until a weighted subspace convergence criterion is met.

We demonstrate the approach with two numerical examples. First is an elliptic

PDE inverse problem, based on a simple model of subsurface flow. Though the dimension of the parameter space in our experiments ranges from 1200 to 10800, the dimension of the LIS remains only around 20 and is empirically discretization-invariant. Exploring the LIS by MCMC and analytically treating the Gaussian complement produces mean and variance fields very similar to those computed via MCMC in the full space. Yet the mixing properties and the computational cost of MCMC in the LIS are dramatically improved over those of full-space MCMC. Our second demonstration is an atmospheric remote sensing problem, where the goal is to infer the concentrations of chemical species in the atmosphere using star occultation measurements, as on the satellite-borne GOMOS instrument. The dimension of the full problem used here was 200 (four gaseous species and 50 altitudes for each), while the dimension of the LIS was 22. Again, dimension reduction significantly improves the mixing properties of MCMC without sacrificing accuracy.

To conclude, our dimension reduction approach appears to offer an efficient way to probe and exploit the structure of nonlinear inverse problems in order to perform Bayesian inference at a large scale, where standard algorithms are plagued by the curse of dimensionality. The approach also opens up interesting further research questions: it may be useful, for instance, to apply reduced-order and surrogate modeling techniques in the LIS, making them applicable to much larger problems than before.

## Acknowledgments

We acknowledge Marko Laine and Johanna Tamminen from the Finnish Meteorological Institute for providing us with the GOMOS figure and codes that served as the baseline for our implementation for the remote sensing example. This work was supported by the US Department of Energy, Office of Advanced Scientific Computing (ASCR), under grant numbers DE-SC0003908 and DE-SC0009297.

## References

- [1] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial Mathematics, Philadelphia, 2005.
- [2] J. P. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*, volume 160. Springer, New York, 2004.
- [3] A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.
- [4] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in practice*, volume 2. CRC press, 1996.
- [5] J. S. Liu. *Monte Carlo strategies in Scientific Computing*. Springer, New York, 2001.
- [6] S. Brooks, A. Gelman, G. Jones, and X. L. Meng, editors. *Handbook of Markov Chain Monte Carlo*. Taylor & Francis, 2011.
- [7] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7:110–120, 1997.
- [8] G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60:255–268, 1998.
- [9] G. O. Roberts and J. S. Rosenthal. Optimal scaling of various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.
- [10] J. C. Mattingly, N. Pillai, and A. M. Stuart. Diffusion limits of the random walk Metropolis algorithm in high dimensions. *Annals of Applied Probability*, 22:881–930, 2012.
- [11] N. S. Pillai, A. M. Stuart, and A. H. Thiery. Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions,. *Annals of Applied Probability*, 22:2320–2356, 2012.

- [12] A. Beskos, G. O. Roberts, A. M. Stuart, and J. Voss. MCMC methods for diffusion bridges. *Stochastic Dynamics*, 8(3):319–350, 2008.
- [13] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28:424–446, 2013.
- [14] H. P. Flath, L. C. Wilcox, V. Akcelik, J. Hill, B. van Bloemen Waanders, and O. Ghattas. Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial hessian approximations. *SIAM Journal on Scientific Computing*, 33(1):407–432, 2011.
- [15] A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio, and Y. Marzouk. Optimal low-rank approximation of linear bayesian inverse problems. *Preprint*, 2014.
- [16] Y. M. Marzouk and H. N. Najm. Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. *Journal of Computational Physics*, 228:1862–1902, 2009.
- [17] K. Karhunen. Über lineare methoden in der wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys.*, 37:1–79, 1947.
- [18] M. Loève. *Probability theory, Vol. II*, volume 46 of *Graduate Texts in Mathematics*. Springer-Verlag, Berlin, 4 edition, 1978.
- [19] J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3):A1460–A1487, 2012.
- [20] Noemi Petra, James Martin, Georg Stadler, and Omar Ghattas. A computational framework for infinite-dimensional Bayesian inverse problems: Part II. stochastic Newton MCMC with application to ice sheet flow inverse problems. *Preprint*, 2013. arXiv:1308.6221.
- [21] M. Lassas, E. Saksman, and S. Siltanen. Discretization invariant Bayesian inversion and Besov space priors. *Inverse Problems and Imaging*, 3(1):87–122, 2009.
- [22] W. Förstner and M. Boudewijn. A metric for covariance matrices. In *Quo vadis geodesia*, pages 113–128. University Stuttgart, 1999.
- [23] G. H. Golub and C. F. Van Loan. *Matrix Computations*. JHU Press, 2012.
- [24] N. Halko, P. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.
- [25] E. Liberty, F. Woolfe, P. G. Martinsson, V. Rokhlin, and M. Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of the National Academy of Sciences*, 104(51):20167–20172, 2007.
- [26] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.
- [27] C. Andrieu and E. Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *The Annals of Applied Probability*, 16(3):1462–1505, 2006.
- [28] Y. F. Atchade. An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift. *Methodology and Computing in Applied Probability*, 8(2):235–254, 2006.
- [29] G. O. Roberts and J. S. Rosenthal. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability*, 44(2):458–475, 2007.
- [30] G. O. Roberts and J. S. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.
- [31] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.
- [32] G. Casella and C. P. Robert. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.
- [33] F. Li, Q. Dai, W. Xu, and G. Er. Weighted subspace distance and its applications to object recognition and retrieval with image sets. *Signal Processing Letters, IEEE*, 16(3):227–230, 2009.
- [34] T. Cui, K. J. H. Law, and Y. M. Marzouk. Dimension-independent likelihood-informed MCMC. *arXiv preprint*, 2014.
- [35] H. Haario, M. Laine, M. Lehtinen, E. Saksman, and J. Tamminen. Markov chain Monte Carlo methods for high dimensional inversion in remote sensing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66:591–608, 2004.